Frutos-López, M., Mejía-Ocaña, A. B., Sanz-Rodríguez, S., Peláez-Moreno, C., Díaz-de-María, F. & Pizlo, Z. (2012). A simplified subjective video quality assessment method based on Signal Detection Theory. In Domanski, M., et al. (eds.). *2012 Picture Coding Symposium PCS 2012: Proceedings*. (pp. 237-240). IEEE.
DOI: http://dx.doi.org/10.1109/PCS.2012.6213336

A simplified subjective video quality assessment method based on Signal Detection Theory

Manuel de-Frutos-López*, Ana Belén Mejía-Ocaña*, Sergio Sanz-Rodríguez*, Carmen Peláez-Moreno*, Fernando Díaz-de-María * and Zygmunt Pizlo†

* Department of Signal Theory and Communications, University Carlos III Madrid, Spain.

† Department of Psychological Sciences, School of Electrical and Computer Engineering, Purdue University, USA.

***Abstract*—A simplified protocol and associated metrics based on Signal Detection Theory (SDT) for subjective Video Quality Assessment (VQA) is proposed with the aim of filling the gap existing between the lack of discrimination abilities of objective Quality Estimates (specially when perceptually motivated processing methods are involved) and the costly normative subjective quality tests. The proposed protocol employs a reduced number of assessors and provides a quality ranking of the methods being evaluated. It is intended for providing the rapid experimental turn around necessary for developing algorithms. We have validated our proposal by corroborating with our test a well-known result for the video coding community: the quality benefits of including an in-loop deblocking filter. A software interface to design and administrate the test is also made publicly available.**

***Index Terms*—Video Quality Assessment, Subjective Quality, Signal Detection Theory, Statistical Decission Theory, Pair Comparison.**

I. Introduction

Subjective quality assessment is of paramount importance for the characterization of the performance of every media processing or transmission system whose end user is a human being. The main drawback of these methods is always their cost, which limits the number of variants or parameters of the investigated methods that can be explored. Hence, normative protocols as described in [1] are usually too cumbersome to assess the effects of minor algorithmic variations and not feasible for quotidian lab experimentation.

Alternatively, objective Video Quality Assessment (VQA) metrics can be employed for video quality evaluation. Simple metrics such as the Mean Squared Error (MSE) and the Peak Signal to Noise Ratio (PSNR) are frequently used. Other more sophisticated methods for VQA have been instead proposed aiming at modelling functional components of the Human Visual System (HVS). Some well-known algorithms include the Moving Pictures Quality Metric (MPQM), the Sarnoff Just Noticeable Difference (JND) vision model [2], and the Digital Video Quality (DVQ) metric [3]. However, these methods are highly complex and do not adequately account for the temporal distortions due to alterations of the motion trajectories.

A new framework in VQA attempting to measure features that HSV associates with loss of quality has recently grown on popularity. Features such as blocking effect, blur, edge and texture information, etc. are measured in both the reference and distorted video sequences so the discrepancies encountered are indicators of visual quality. Popular VQA algorithms using this approach include the Video Quality Metric (VQM) [4], extensions of the Structural SIMilarity (SSIM) index [5] for video signals [6], and the recently proposed MOtion-based Video INtegrity Evaluation (MOVIE) index [7].

Most of these VQA methods provide a good performance when measuring typical spatial and temporal distortions produced during the acquisition, processing, coding and transmission. Nevertheless, they could result in a lack of effectiveness to evaluate specific perceptual video coding systems where, for instance, the salient areas of visual attention are given a higher bit allocation precedence than the remaining areas.

In this paper we put forward a simplified subjective test within a SDT framework including elements from behavioral sciences for the design of the data acquisition procedure. In particular, the subjective quality comparison of two (or more) alternative methodologies can be obtained following these principles:

- The comparison is made indirectly since only pairs of video sequences generated with the same method (but perhaps using different values of the free parameters) are presented, requiring a simple *yes/no* answer. In the case study of Sec. IV, for example, we employ different bitrate reductions of each of the two contender video coders to obtain such pairs. In this way, we force assessors to focus on the depth of the distortions produced by each alternative algorithm along with the bitrate reductions avoiding the difficulty of comparing two different techniques that are likely to produce distortions of different nature.
- After each trial the assessor is provided with the correct answer since, thanks to the previously mentioned design, the correct answer is known *a priori* (in our example, the reference bitrate is always better than any bitrate reduction). This contributes to a better definition of the decision threshold and is more motivating since assessors always try to guess the rationale behind the experiment.
- Sessions can always be paused and resumed whenever the assessors feel tired.

As a consequence, the following advantages are obtained:

- Small number of assessors are needed though the number of trials per session needs to be high to achieve statistically significant results and therefore the sessions are usually of long duration.
- It complies better with hypothesis testing statistical as-

sumptions since both hypothesis within a pair are expected to have similar pdf shapes.
- As per design, the number of trials are always balanced among pairs simplifying the statistical analysis.
- It is more suitable for detecting small improvements than scale-based quality estimators (e.g. Mean Opinion Score (MOS)) since it only demands from the assessor a simple *yes/no* answer. As we will demonstrate in Sec. IV-B, the assessor's inability to decide in particular situations is correctly acknowledged by the employed measure.

This paper is organized as follows: an introduction to the signal detection framework in section II is followed by our interpretation of this framework for (VQA). A case study to validate our method is presented in section IV. Conclusions and future work close the paper.

II. A Signal Detection Theory Framework

SDT was introduced into the field of psychology and it has been widely adopted as a means of empirical framework design specially in tasks that aim at explaining various aspects of human cognition. The purpose of this paper is to bow for the re-adoption of this theory for VQA. An extensive description of the fundamentals of SDT can be found in [8]. For a comprehensive and more up to date review the reader is referred to [9].

The original goal of SDT is to find out if two different types of stimuli are distinguishable. These stimuli are referred to *signal* and *noise*. Therefore, in perceptual experiments, a human assessor is challenged with the task of deciding whether these stimuli can be told apart or not. A total number of K trials of any of the two stimuli (delivered in random order) are presented to the assessor who has to provide a simple *yes/no* answer to the following question: *Is* $trial_k$ *a signal?*

To answer this question, the assessor needs to perceptually *measure* the stimulus as a function of a hidden decision variable that is only available in his mind since it is the result of a perceptual evaluation. In particular, a *criterion* threshold needs to be placed in a certain position of the decision variable axis. We will adopt the convention (the reversed would be equally admisible) that if the result of the perceptual *measurement* exceeds the criterion, then the assessor will reply *yes* to the SDT question. Otherwise, the answer will be *no*.

The degree to which the difference between *signal* and *noise* can be perceived by the assessor will be inversely proportional to the amount of overlap between their distributions across the decision variable. Unfortunately, this cannot be directly observed and therefore, we need a way to indirectly measure the distance between the two distributions.

The experiments must be designed so that the correct answer to the SDT question is known *a priori* which allows the computation of the following contingency table:

		Stimulus	
		Signal	Noise
Response	'yes' (Signal)	H	FA
	'no' (Noise)	M	CR

where H represents the number of Hits (also known as correct answers, true positives or correct detections), FA is the number of False Alarms (or false positives or incorrect detections), M is the number of Misses (also false negatives or omission errors) and CR is the number of Correct Rejections (or true negatives).

From these values, parametric and non parametric estimations of the separation of the two distributions can be made computing what is called a *sensitivity* index [10]–[13]. A popular parametric estimation is called d' (pronounced *dee-prime*).

For the calculation of d' we apply the simple equation:

$$d' = \Phi^{-1}(HR) - \Phi^{-1}(FAR) \tag{1}$$

where $\Phi^{-1}(p)$ is the *quantile* function or the inverse of the normal cumulative density function with $\sigma = 1$ [1]:

$$\Phi^{-1}(p) \equiv z_p = \sqrt{2}\,\mathrm{erf}^{-1}(2p-1) \quad p \in (0,1) \tag{2}$$

being $\mathrm{erf}(x)$ the Gauss error function, HR the Hit Rate and FAR, the False Alarm Rate. See details in [14].

It is now commonly accepted that there is very small variability among human observers when it comes to basic visual functions such as detection and discrimination among stimuli. As a result, it has become a common practice to test only a few subjects (2-5) in a given study, as long as reliable experimental methodology, such as signal detection, is used, which measures the percept unconfounded with response (decision) bias.

III. Signal Detection Theory for VQA

Let us assume that our goal is to compare N video coding alternatives. To apply SDT to each technique, we need to provide a controlled way to produce $(R+1)$ different quality samples from each of the N techniques (at least two) : the first of them will be considered 'the reference' and the remaining R will be tested against it in individual experimental sessions. This can be easily accomplished by producing R reductions of a reference bitrate from each of the coding proposals. In general, every processing technique is likely to adopt some kind of trade-off that could be exploited for these purposes.

Then, $N \times R$ sessions must be designed where the k^{th} trial consists of the presentation of a pair of versions of the same video sequence processed with the n^{th} technique in random order: the reference version and its r^{th} bitrate reduction. Thus, two coding techniques are never directly confronted since our goal is to perform an indirect comparison by quantifying for which techniques the bitrate reductions are more noticeable.

It is worth noting that in order to compute HR and FAR the best quality sample of the pair must be known *a priori*. This requirement is easily achieved in our case study by assuming that, for a given video coder, higher bitrates imply a better quality. Though this can be potentially controversial,

[1]Therefore, the resulting d' will be proportional to the $\sigma_s = \sigma_n$ or, in other words, d' is measured in standard deviation units which, for the purposes of this paper is irrelevant.

our opinion is that this prerequisite can be easily met in a variety of situations since trade-offs (as the bitrate-quality of video coders) abound in media processing.

Then the SDT question needs to be recast as:

VQA question: *Is the first video sequence of* $trial_k(n, r)$ *better in comparison with the second?* [2]

Now, the assessors need to answer based on a criterion they set on a hidden decision variable. The *signal* and *noise* distributions in section II are now, respectively, the distribution of trials where *the first sequence is better than the second* (stimulus $S1$ in the sequel) and the distribution of trials where *the second sequence is better than the first* (stimulus $S2$). Collecting the answers as in the table of section II allows the calculus of d' as defined in equation (1).

IV. A CASE STUDY: THE DEBLOCKING FILTER IN H.264

To validate our method we have re-evaluated a very well-known result in video coding: the inclusion of an in-loop deblocking filter produces perceptually enhanced quality [16].

A. Experimental setup

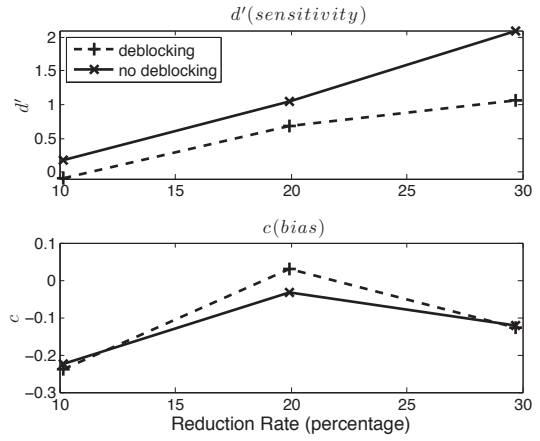
A simplified version of the Pair Comparison method described in [1], recommended for its high discriminatory ability when the test items are almost identical in quality was carried out. Specifically, those recommendations for subjects and sequences selection were followed so that parameters of the experimental setup laid between recommended margins.

A set of 90 clips, consisting mostly of sections of traditional test video sequences, were arranged for the experiment. The duration of each clip varies between 3 and 7 seconds. A reference bit rate of 512 kbps was selected for the experiments as well as three different bit rate reductions for each base bit rate. All these operating points were accurately achieved by means of the variable bit rate control algorithm described in [17]. implemented on the Joint Video Team (JVT) H.264 reference software version JM 12.2. As mentioned before, two different versions of this encoder were employed in which the only difference is the activation/deactivation of the in-loop deblocking filter.

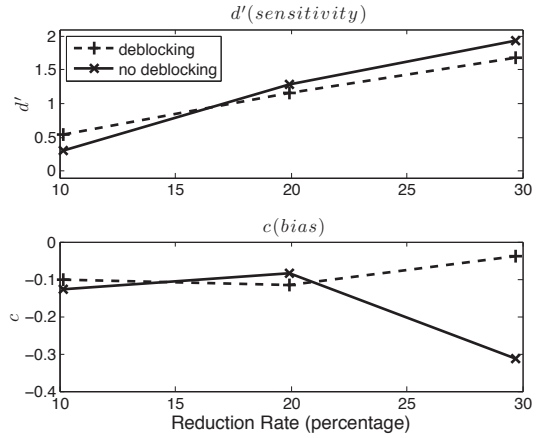
For each one of the 6 test sessions, corresponding to three bit Rate Reductions (RR) for both encoder versions, test clips were shown in pairs of reference and reduced-rate clip versions, and the aim of the test was to measure the ability of the assessors to detect the reduced bit rate version of each pair, which was shown twice in reverse order to cancel the bias effect, i.e. the tendency of each subject to choose the first or second sequence as the best when in doubt. For each pair, the correct answer was fed back to the assessor after his choice in order to train the subject. Therefore, 180 trials were evaluated altogether in each of the sessions by each of the 4 assessors, two of them naive and two experts.

[2]Of course, this question can be rephrased into a more natural one as 'Which of the two sequences is better?' that is what we actually implement in the interface available on-line [15]. However, from a theoretical point of view, we adopt the more artificial question that allows for a *yes/no* answer.

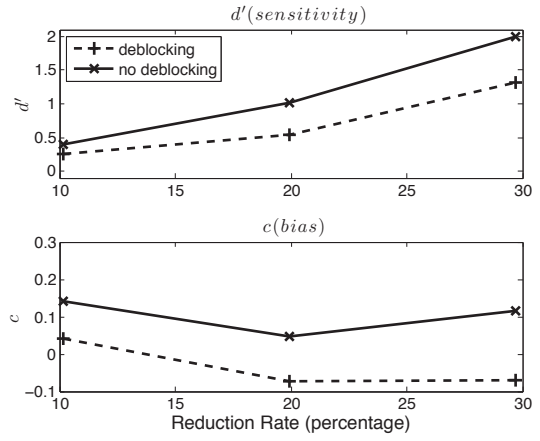
B. Results



(a)



(b)



(c)

Fig. 1. Sensitivity (d') and Bias (c) to Rate Reductions of 10, 20 and 30% of the video coder with and without the deblocking filter for assessor $\#1 - \#3$ (resp. (a) - (c))

The results obtained for assessors $\#1$, $\#2$ and $\#3$ are depicted in figure 1. The fourth assessor's results are not presented here since the sensitivity indexes collected where too low ($d' < 0.3$ for all the sessions), indicating either a lack

of attention or some kind of visual impairment. In our opinion, far from considering this a failure we regard it as a desirable feature since the results clearly alert of these kind of situations, preventing misuses. Biases' metrics are also included here for the sake of completeness.

The first obvious observation that can be made from the positive slope of the sensitivity lines is that, as the RR increases it is easier to distinguish between the reference and the reduced bit rate clips. Thus, for a RR of 30% and for all the assessors, d' is around 1.5 when the deblocking filter is on and near 2.0 when it is absent. From this, we can conclude that the distributions of stimulus S1 and S2 exhibit only a small overlapping area. On the other end of the plots, when the RR is only a 10%, d' is around 0.0, indicating that the distributions of both stimulus almost totally overlapped.

More importantly, and with the only exception of the RR=10% and assessor $\#2$ (figure 1(b)), the *no deblocking* sensitivity lines are always on top of the *deblocking* ones. As we already said, this is a very well-known conclusion that corroborates that our quality assessment method is working properly since it clearly indicates that using the in-loop deblocking filter of H.264 improves the quality of the resulting coded sequences or, in other words, when this filter is absent the bitrate reductions are more noticeable.

It is worth noting that the aforementioned exception is not significant since it only appears when the RR is very low and d' values are, accordingly, very low as well. In essence, if the bit rates of two samples of the same sequence are very close, it is not possible to tell them apart and, therefore, it does not matter if the deblocking filter is on or not. On the other end, for RR of 30% the differences are systematic for all three observers. We have performed simple z tests on each of the pairs compared that demonstrate that the difference between applying or not the deblocking filter is statistically significant with the exception of assessor $\#2$.

V. Conclusions

In this paper we have proposed a novel subjective quality test for rapid experimentation turn-around based on SDT where:

- The subjective quality measure provided is always *relative* as the presentation of the trials is based on a pair comparison. The results of the assessments are rankings of methods.
- The assessors are required to answer an elementary *yes/no* question avoiding the difficulty of gradings. This makes our proposal suitable for discerning among similar methods by making the decision simpler and more natural.
- Two processing techniques are never directly confronted since our goal is to perform an indirect comparison by quantifying the influence of their trade-off parameters.
- The number of assessors can be small as it is only required for validate the results of the ranking.

We have validated our proposal by testing it on the well-known result that the deblocking filter improves the perceptual quality in video coding.

Our future lines of work contemplate the extension to threshold analysis by using adaptive psychophysical procedures like PEST or QUEST.

Acknowledgment

This work has been partially supported by the regional project CCG10-UC3M/TIC-5304 (Comunidad Autónoma de Madrid - UC3M) and by National Grant TEC2011-26807 of the Spanish Ministry of Science and Innovation.

References

[1] ITU-T Study Group 9, "ITU-T Rec. P.910 (04/2008) subjective video quality assessment methods for multimedia applications," *Series P: Telephone transmission quality, telephone installations, local line networks - Audiovisual quality in multimedia services*, pp. 1–42, Feb 2009.

[2] J. Lubin, "Digital images and human vision," A. B. Watson, Ed. Cambridge, MA, USA: MIT Press, 1993, ch. The use of psychophysical data and models in the analysis of display system performance, pp. 163–178.

[3] A. B. Watson, J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *Journal of Electronic Imaging*, vol. 10, no. 1, p. 20, 2001.

[4] M. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *Broadcasting, IEEE Transactions on*, vol. 50, no. 3, pp. 312 – 322, 2004.

[5] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *Image Processing, IEEE Transactions on*, vol. 13, no. 4, pp. 600 –612, 2004.

[6] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *J. Opt. Soc. Am. A*, vol. 24, no. 12, pp. B61–B69, Dec 2007.

[7] K. Seshadrinathan and A. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *Image Processing, IEEE Transactions on*, vol. 19, no. 2, pp. 335 –350, 2010.

[8] C. W. Helstrom, *Statistical theory of signal detection*. Pergamon Press, 1960.

[9] H. Stanislaw and N. Todorov, "Calculation of signal detection theory measures," *Behavior Research Methods Instruments and Computers*, vol. 31, pp. 137–149, 1999.

[10] A. L. Brophy, "Alternatives to a table of criterion values in signal detection theory," *Behavior Research Methods, Instruments and Computers*, vol. 18, no. 3, pp. 285–286, Dec 1986.

[11] W. Tanner and J. Swets, "A decision-making theory of visual detection," *Psychol Rev*, vol. 61, no. 6, pp. 401–409, 1954.

[12] W. Donaldson, "Measuring recognition memory," *Journal of Experimental Psychology: General*, vol. 121, no. 3, pp. 275–277, 1992.

[13] K. Feenan and J. Snodgrass, "The effect of context on discrimination and bias in recognition memory for pictures and words." *Memory & cognition*, vol. 18, no. 5, p. 515, 1990.

[14] M. de Frutos-López, A. B. Mejía-Ocaña, S. Sanz-Rodríguez, C. Peláez-Moreno, F. Díaz-de-María, and Z. Pizlo, "A simplified subjective video quality assessment method for rapid experimental turn-around," *IEEE Transactions on Circuits and Systems for Video Technology*, 2012 (submitted).

[15] G. P. M. Multimedia Processing Group University Carlos III Madrid. (2011, Dec.) A simplified subjective video quality assessment method java interface. [Online]. Available: gpm.tsc.uc3m.es

[16] P. List, A. Joch, J. Lainema, G. Bjontegaard, and M. Karczewicz, "Adaptive deblocking filter," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 13, no. 7, pp. 614–619, 2003.

[17] M. de Frutos-Lopez, O. del Ama-Esteban, S. Sanz-Rodriguez, and F. Diaz-de Maria, "A two-level sliding-window vbr controller for real-time hierarchical video coding," in *Image Processing (ICIP), 2010 17th IEEE International Conference on*, 2010, pp. 4217–4220.